Bayesian inference:
Bayes's theorem and beyond

Tom Loredo
Cornell Center for Astrophysics and Planetary Science,
Carl Sagan Institute,
& Dept. of Statistics and Data Science, Cornell U.
http://hosting.astro.cornell.edu/~loredo/

ASTRO 4523 — 18 Sep 2025

Lecture resources

- SummerSchool2025-IntroBayes (GitHub)
  - ▶ These slides
  - ▶ PSU Center for Astrostatistics Summer School slides and recorded lectures ($\sim$ 3 hr)
  - ▶ Passcode for lecture videos: `Bayes25SlidesAccess!`
- Recent "second tutorial" for relative newcomers somewhat familiar with Bayes's theorem:
  - ▶ Bayesian inference: more than Bayes's theorem (ADS) (use journal version, not arXiv version)

The "classical" understanding of probability

Pierre Simon Laplace (1819)

Probability theory is nothing but *common sense reduced to calculation*.

James Clerk Maxwell (1850)

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic, but the actual science of *Logic is conversant at present only with things either certain, impossible, or entirely doubtful*, none of which (fortunately) we have to reason on. Therefore *the true logic of this world is the calculus of Probabilities*, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

Harold Jeffreys (1931)

If we like there is no harm in saying that a *probability expresses a degree of reasonable belief*. . . . 'Degree of confirmation' has been used by Carnap, and possibly avoids some confusion. But whatever verbal expression we use to try to convey the primitive idea, this expression cannot amount to a definition. *Essentially the notion can only be described by reference to instances where it is used*. It is intended to express *a kind of relation between data and consequence* that habitually arises in science and in everyday life, and the reader should be able to recognize the relation from examples of the circumstances when it arises.

Lies, damn lies, and. . .

Benjamin Disraeli

"There are three kinds of lies: lies, damned lies, and statistics."

Lies, damn lies, and. . .

Benjamin Disraeli

Mark Twain (1906)

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark *attributed to* Disraeli would often apply with justice and force:

"There are three kinds of lies: lies, damned lies, and statistics."

Lies, damn lies, and. . .

Benjamin Disraeli

Mark Twain (1906)

> Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark *attributed to* Disraeli would often apply with justice and force:
>
> *"There are three kinds of lies: lies, damned lies, and statistics."*

Lie? *That Disraeli ever said or wrote it!*

Lies, damn lies, and. . .

Benjamin Disraeli

Mark Twain (1906)

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark *attributed to* Disraeli would often apply with justice and force:

"There are three kinds of lies: lies, damned lies, and statistics."

Lie? *That Disraeli ever said or wrote it!*

Damn lie? *That the statement was originally about statistics!*

Liars, damn liars, and. . .

Sir Robert Giffen (1892)

> An old jest runs to the effect that there are three kinds of comparison among liars. There are liars, there are outrageous liars, and there are *scientific experts*.

> This has lately been adapted to throw dirt upon statistics. There are three degrees of comparisons, it is said, in lying. There are lies, there are outrageous lies, and there are statistics.

Liars, damn liars, and. . .

Sir Robert Giffen (1892)

> An old jest runs to the effect that there are three kinds of comparison among liars. There are liars, there are outrageous liars, and there are *scientific experts*.
>
> This has lately been adapted to throw dirt upon statistics. There are three degrees of comparisons, it is said, in lying. There are lies, there are outrageous lies, and there are statistics.
>
> Statisticians can afford to laugh at and profit by jokes at their expense. There is so much knowledge which is unobtainable except by statistics. . .

"On international statistical comparisons," *Economic Journal* (1892)

See http://www.york.ac.uk/depts/maths/histstat/lies.htm

Scientific method

Science is more than a body of knowledge; it is a way of thinking.
The method of science, as stodgy and grumpy as it may seem,
is far more important than the findings of science.
—Carl Sagan

Scientists *argue!*

Argument ≡ Collection of statements comprising an act of reasoning from *premises* to a *conclusion*

A key goal of science: Explain or predict *quantitative measurements* (data!)

Data analysis: Constructing and appraising arguments that reason from data to interesting scientific conclusions (explanations, predictions)

Truths, subtle truths. . . and statistical inference

Science is more than a body of knowledge; it is a way of thinking.
The method of science, as stodgy and grumpy as it may seem,
is far more important than the findings of science.
                                    —Carl Sagan

Bayesian viewpoint

- Key part of scientific method: Arguing for/against models & theories based on *quantitative data*

- Statistical inference aims to *quantify the strength of data-based arguments*

- Bayesian (probable) inference aims to do this particularly directly:

  Calculate p(Hypothesis | Data)

  (including for *composite* hypotheses—when there are different ways the hypothesis may be true)

Agenda

❶ Quantifying uncertainty with probability

❷ Nuisance parameters and marginalization

❸ Model comparison and marginal likelihood

❹ (Measurement errors and latent variable marginalization)

❺ Recap/takeaways

Agenda

❶ Quantifying uncertainty with probability

❷ Nuisance parameters and marginalization

❸ Model comparison and marginal likelihood

❹ (Measurement errors and latent variable marginalization)

❺ Recap/takeaways

Probability theory as (generalized) logic

"Logic can be defined as *the analysis and appraisal of arguments*"
—Gensler, *Intro to Logic*

Build arguments with *propositions* and logical
operators/connectives

- *Propositions:* Statements that may be true or false

    $\mathcal{P}$: Universe can be modeled with ΛCDM

    A : $\Omega_{\text{tot}} \in [0.9, 1.1]$

    B : Ω_Λ is not 0

    $\overline{B}$: "not B," i.e., $\Omega_\Lambda = 0$

    Events in freq. PT are propositions about outcomes in
    repeated trials

- *Connectives:*

    $A \wedge B$ or A, B : A and B are *both* true

    $A \vee B$: A or B is true, or both are

Arguments

Argument: Assertion that an *hypothesized conclusion*, H, follows from *premises*, $\mathcal{P} = \{A, B, C, \ldots\}$ (take "," = "and")

Notation:

$H|\mathcal{P}$:   Premises $\mathcal{P}$ imply H

H may be deduced from $\mathcal{P}$

H follows from $\mathcal{P}$

H is true given that $\mathcal{P}$ is true

Deductive logic applies when we can *reason with certainty*; can model this with *Boolean algebra* over $\{0, 1\}$ (False, True)

Classical/Bayesian PT applies when we must *reason amidst uncertainty*; it quantifies degree of certainty on a $[0, 1]$ scale, providing a mathematical model for *inductive reasoning*

Probability as argument strength

$$P(H|\mathcal{P}) \equiv \text{strength of argument } H|\mathcal{P}$$

$$
\begin{aligned}
P &= 1 \; \rightarrow \; \text{Argument is } \textit{deductively valid} \\
&= 0 \; \rightarrow \; \text{Premises imply } \overline{H} \\
&\in (0,1) \; \rightarrow \; \text{Degree of deducibility/entailment}
\end{aligned}
$$

Mathematical model for inductive reasoning

$$
\begin{aligned}
\text{'AND' (product rule):} \quad P(A \wedge B|\mathcal{P}) &= P(A|\mathcal{P})\,P(B|A \wedge \mathcal{P}) \\
&= P(B|\mathcal{P})\,P(A|B \wedge \mathcal{P})
\end{aligned}
$$

$$
\begin{aligned}
\text{'OR' (sum rule):} \quad P(A \vee B|\mathcal{P}) &= P(A|\mathcal{P}) + P(B|\mathcal{P}) \\
&\quad - P(A \wedge B|\mathcal{P})
\end{aligned}
$$

$$
\text{'NOT':} \quad P(\overline{A}|\mathcal{P}) = 1 - P(A|\mathcal{P})
$$

Interpreting probability distributions
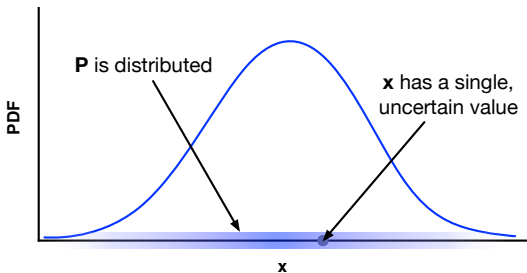
Frequentist

Probabilities are always (limiting) rates/proportions/frequencies that *quantify variability* in a sequence of trials. $p(x)$ describes how the *values of x* would be distributed among *infinitely many trials*:
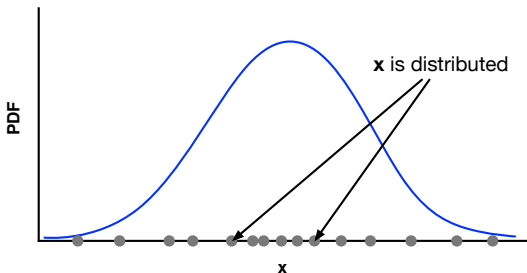
x must be a quantity exhibiting variability across replications or a population—data values, but not parameter values (hypothesis labels)

Bayesian

Probability *quantifies uncertainty* in an inductive inference. $p(x)$
describes how *probability* is distributed over the possible values x
might have taken in *the single case before us*:

x may denote any quantity we may be uncertain of—data values, or
parameter values

Frequentist vs. Bayesian statements

"The data D_{obs} support hypothesis H . . . "

Frequentist assessment

"H was selected with a procedure that's right 95% of the time over a set $\{D_{hyp}\}$ that includes D_{obs}."

Probabilities are properties of *procedures*, not of particular results. Guaranteed long-run performance is the *sine qua non*.

Bayesian assessment

"The strength of the chain of reasoning from the model and D_{obs} to H is 0.95, on a scale where $1 =$ certainty."

Probabilities are associated with arguments that directly refer only to *actually observed data*.

Long-run performance must be separately evaluated (and is typically good by frequentist criteria).

Aside: Bayesian lingo

- *Bayesian probability theory*: Assigning and manipulating probabilities interpreted as argument strength/degree of belief/entailment strength

- *Bayesian inference*: Computing posterior probabilities for hypotheses (e.g., statements about parameters), or posterior predictive probabilities for future data

- *Bayesian decision theory*: Making optimal decisions given data, accounting for consequences of good or bad decisions (utility or loss functions); includes experimental design

- *Bayesian data analysis (BDA)*: Any mode of data analysis using a Bayesian interpretation of probability, including inference & decision/design, but also prediction, model checking (predictive tests), EDA/visualization, data reduction. . .

- *Bayesian statistics*: Informal catch-all for any one or more of the above

- *Bayesian workflow*: Emerging term of art for best practices, including model checking and adjustment/refinement

*We'll discuss only Bayesian inference here,
but the other areas are also important*

Ten Great Ideas about Chance (2017)

Semi-technical survey by a leading statistician/mathematician (Diaconis) and a leading philosopher of science (Skyrms). "This is a history book, a probability book, and a philosophy book."

"To anyone with an interest in probability or statistics, this is a book you must read. . . . [It] is far-ranging and can be read at many levels, from the novice to the expert. It is also thoroughly engaging." —David M. Bressoud, UMAP Journal

The Bayesian inference recipe

Assess hypotheses by calculating their probabilities $p(H_i| \ldots)$ conditional on known and/or presumed information (including observed data) using the rules of probability theory.

Probability Theory Axioms

$$\mathcal{C} \equiv \textit{context}, \text{ initial set of premises}$$

$$\text{'OR' (sum rule):} \quad P(H_1 \vee H_2 | \mathcal{C}) = \begin{aligned} &P(H_1|\mathcal{C}) + P(H_2|\mathcal{C}) \\ &- P(H_1, H_2|\mathcal{C}) \end{aligned}$$

$$\text{'AND' (product rule):} \quad \begin{aligned} P(H_i, D_{\mathrm{obs}} | \mathcal{C}) &= P(H_i|\mathcal{C})\, P(D_{\mathrm{obs}}|H_i, \mathcal{C}) \\ &= P(D_{\mathrm{obs}}|\mathcal{C})\, P(H_i|D_{\mathrm{obs}}, \mathcal{C}) \end{aligned}$$

$$\text{'NOT':} \quad P(\overline{H_i}|\mathcal{C}) = 1 - P(H_i|\mathcal{C})$$

Two Important Theorems

Bayes's Theorem (BT)

Consider the *joint probability* for a hypothesis and the observed data, $P(H_i, D_{\mathrm{obs}}|\mathcal{C})$, using the product rule:

$$
\begin{aligned}
P(H_i, D_{\mathrm{obs}}|\mathcal{C}) &= P(H_i|\mathcal{C})\, P(D_{\mathrm{obs}}|H_i, \mathcal{C}) \\
&= P(D_{\mathrm{obs}}|\mathcal{C})\, P(H_i|D_{\mathrm{obs}}, \mathcal{C})
\end{aligned}
$$

Solve for the *posterior probability* for H_i (adds a premise!):

$$
P(H_i|D_{\mathrm{obs}}, \mathcal{C}) = \frac{P(H_i, D_{\mathrm{obs}}|\mathcal{C})}{P(D_{\mathrm{obs}}|\mathcal{C})} = P(H_i|\mathcal{C})\, \frac{P(D_{\mathrm{obs}}|H_i, \mathcal{C})}{P(D_{\mathrm{obs}}|\mathcal{C})}
$$

Theorem holds for any propositions, but for hypotheses & data the factors have names:

posterior $\propto$ prior $\times$ likelihood
(all "for H_i")

norm. const. $P(D_{\mathrm{obs}}|\mathcal{C}) = $ *prior predictive* for D_{obs}

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ ("suite;" $\mathcal{C}$ asserts one of them must be true),

$$\begin{aligned}
\sum_i P(A, B_i|\mathcal{C}) &= \sum_i P(B_i|A, \mathcal{C})P(A|\mathcal{C}) = P(A|\mathcal{C}) \\
&= \sum_i P(B_i|\mathcal{C})P(A|B_i, \mathcal{C})
\end{aligned}$$

If we do not see how to get $P(A|\mathcal{C})$ directly, we can find a set $\{B_i\}$ and use it as a "basis"—*extend the conversation*:

$$P(A|\mathcal{C}) = \sum_i P(B_i|\mathcal{C})P(A|B_i, \mathcal{C})$$

If our problem already has B_i in it, we can use LTP to get $P(A|\mathcal{P})$ from the joint probabilities—*marginalization*:

$$P(A|\mathcal{C}) = \sum_i P(A, B_i|\mathcal{C})$$

Joseph Blitzstein (Harvard statistician) on LTP (paraphrased):

> In most areas of math, when you're stuck, saying, "I wish I knew this or that" doesn't help you. In probability theory, saying "I wish I knew this" suggests what to condition on; then you condition on it, compute *as if* you knew it, and then average over those possibilities.
>
> *I didn't name the law of total probability, but if I had, I would have just called it* **wishful thinking**.
>
> — YouTube lecture on conditional probability (15:48)

LTP example 1: Take $\mathcal{C}$ to specify fair roll of a die, $A =$ "An even number comes up," $B_i =$ "face i comes up" ($i = 1$ to 6)

$$
\begin{aligned}
P(A|\mathcal{C}) &= \sum_{i=1}^{6} P(A, B_i|\mathcal{C}) \\
&= \sum_{i=1}^{6} P(B_i|\mathcal{C})P(A|B_i,\mathcal{C}) \\
&= \frac{1}{6} \times (0 + 1 + 0 + 1 + 0 + 1) = \frac{1}{2}
\end{aligned}
$$

LTP example 2: With context $\mathcal{C}$, take $A = D_{\mathrm{obs}}$, $B_i = H_i$; then

$$
\begin{aligned}
P(D_{\mathrm{obs}}|\mathcal{C}) &= \sum_i P(D_{\mathrm{obs}}, H_i|\mathcal{C}) \\
&= \sum_i P(H_i|\mathcal{C})P(D_{\mathrm{obs}}|H_i,\mathcal{C})
\end{aligned}
$$

prior predictive for D_{obs} = Average likelihood for H_i
(a.k.a. *marginal likelihood*)

Sampling distributions $\rightarrow$ likelihood function

The *sampling distribution* is the probability distribution *for the data*. For a parametric model, we have a *collection* of probability distributions, a bivariate function of hypothetical values of the parameter, and hypothetical values of the data:

$$p(D|\theta) = f(D; \theta)$$

"The" sampling dist'n is the one with $\theta = \theta_*$ (the true value)

The *likelihood function* **fixes** $D = D_{\text{obs}}$, yielding a function only of θ:

$$\mathcal{L}(\theta) = f(D_{\text{obs}}; \theta)$$

This is *not* a probability distribution for θ (θ is on the wrong side of the solidus)

The *posterior distribution* is also a function only of θ:

$$p(\theta|D_{\text{obs}}) \propto \pi(\theta) \times \mathcal{L}(\theta)$$

But we now have a *probability distribution over θ*

Sampling distributions and likelihood function

Sampling distributions **Likelihood function**
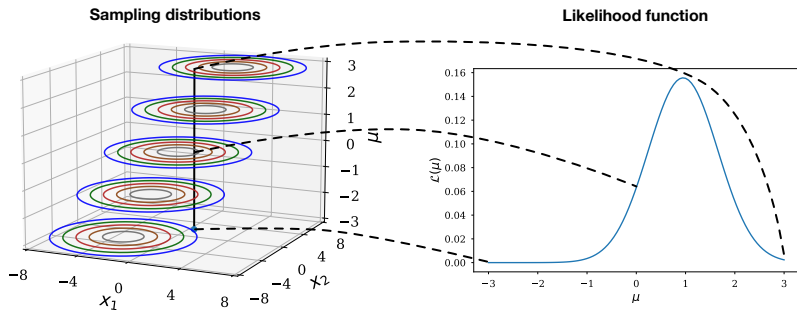
The likelihood function shows how well each of the candidate sampling distributions—labeled by the parameter, μ—predicts the observed data (x_1, x_2); "predictive" or "prognostic" might have been better

The likelihood for the parameter is the (sampling) probability for the observed data; "likelihood for the data" is incorrect usage—it entirely misses the point of likelihood!

Fisher on likelihood

"If we need a word to characterise this relative property of different values of *p*, I suggest that we may speak without confusion of the likelihood of one value of *p* being thrice the likelihood of another, bearing always in mind that *likelihood is not here used loosely as a synonym of probability*, but simply to express the relative frequencies with which such values of the hypothetical quantity *p* would in fact yield the observed sample." (Fisher 1922)

"Likelihood also *differs from probability* in that it is a differential element, and is *incapable of being integrated*: it is assigned to a particular point of the range of variation, not to a particular element [interval]." (Fisher 1922)

"...the integration with respect to *m* is illegitimate and has no definite meaning..." (Fisher 1912)

Likelihood function $\rightarrow$ posterior distribution

The *posterior distribution* is also a function only of θ:

$$p(\theta|D_{\mathrm{obs}}) \propto \pi(\theta) \times \mathcal{L}(\theta)$$

We now have a *probability distribution over θ*

This is just the *starting point*—we now have to do calculations, using all of probability theory to get answers to our questions from the posterior

We'll find ourselves using the *law of total probability* over and over again—**marginalization** (summing/integrating probabilities)

Integrating over parameter space is the key feature distinguishing Bayesian from frequentist data analysis (frequentist methods typically **optimize** over parameter space)

Agenda

Nuisance parameters and marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

Example

We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal s and a background b.

We have additional data just about b.

What do the data tell us about s?

Marginal posterior distribution

To summarize implications for s, accounting for b uncertainty, *marginalize*:

$$
\begin{aligned}
p(s|D, M) &= \int db\ p(s, b|D, M) \\
&\propto p(s|M) \int db\ p(b|s, M)\ \mathcal{L}(s, b) \\
&= p(s|M)\mathcal{L}_m(s)
\end{aligned}
$$

with $\mathcal{L}_m(s)$ the *marginal likelihood function for s*:
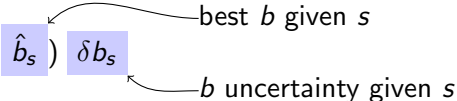
$$
\mathcal{L}_m(s) \equiv \int db\ p(b|s)\ \mathcal{L}(s, b)
$$

Maximum likelihood suggests instead computing the *profile likelihood*:

$$
\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s), \qquad \hat{b}_s = \text{best } b \text{ given } s
$$

Marginalization vs. optimization

For insight: Suppose the prior is broad compared to the likelihood $\rightarrow$ for a fixed s, we can accurately estimate b with max likelihood $\hat{b}_s$, with small uncertainty δb_s.

$$
\begin{aligned}
\mathcal{L}_m(s) &\equiv \int db \, p(b|s) \, \mathcal{L}(s, b) \\
&\approx \quad p(\hat{b}_s|s) \, \mathcal{L}(s, \hat{b}_s) \; \delta b_s
\end{aligned}
$$

best b given s

b uncertainty given s

Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s)$ gets weighted by a *parameter space volume factor*

E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}$, $\quad \sigma_s^2 = \sigma_r^2 + \sigma_b^2$, and δb_s is *const.*

Background *subtraction* is a special case of background *marginalization*.

Flared/skewed/bannana-shaped: $\mathcal{L}_m$ and $\mathcal{L}_p$ *differ*
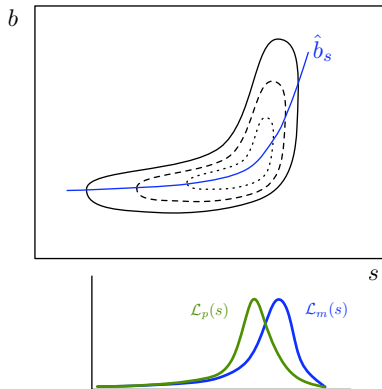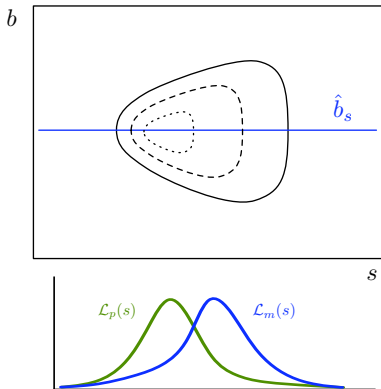
General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$.
Otherwise, they will likely *differ*.

In *measurement error problems* the difference can have dramatic consequences (due to proliferation of latent parameters)

Agenda

Model comparison

Problem statement

$C = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models.

$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model

$$p(M_i|D,C) = p(M_i|C)\frac{p(D|M_i,C)}{p(D|C)}$$

$$\propto p(M_i|C)\mathcal{L}(M_i)$$

$$\mathcal{L}(M_i) \equiv p(D|M_i) = \int d\theta_i \, p(\theta_i|M_i)p(D|\theta_i, M_i)$$

Likelihood for model = Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

Varied terminology: Prior predictive = *Marginal likelihood* = Average likelihood = Global likelihood = (Weight of) Evidence for model

Odds and Bayes factors

A ratio of probabilities for two propositions using the same premises is called the *odds* favoring one over the other:

$$
\begin{aligned}
O_{ij} &\equiv \frac{p(M_i|D,\mathcal{C})}{p(M_j|D,\mathcal{C})} \\
&= \frac{p(M_i|\mathcal{C})}{p(M_j|\mathcal{C})} \times \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})}
\end{aligned}
$$

The data-dependent part is called the *Bayes factor*:

$$
B_{ij} \equiv \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})}
$$

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are marginal/average likelihoods for *composite hypotheses*

An automatic Ockham's razor

Consider *nested models*:

- Simpler model M_1 with parameters θ_1
- "Larger" rival M_2 with parameters $\theta_2 = (\theta_1, \eta)$

$\Rightarrow \mathcal{L}(\hat{\theta}_2) \geq \mathcal{L}(\hat{\theta}_1)$ so *maximum likelihood* can never favor M_1

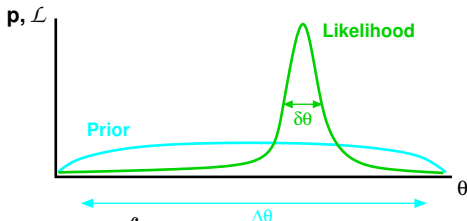But what about $p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i)$?

Prior predictive distributions

Normalization implies *there must be data that favor M_1*:
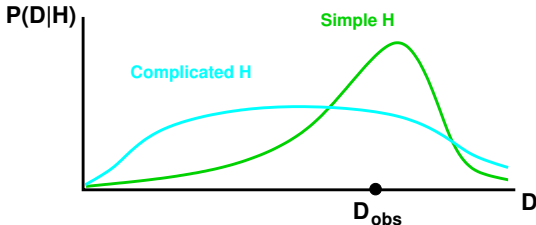
The Ockham factor

$$p(D|M_i) = \int d\theta_i \, p(\theta_i|M) \, \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i$$

$$\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i}$$

$$= \text{Maximum Likelihood} \times \text{Ockham Factor}$$

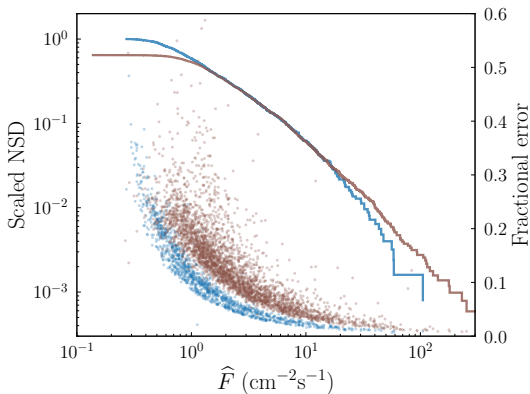Models with more parameters often make the data more probable — *for the best fit*

Ockham factor penalizes models for "wasted" volume of parameter space

Quantifies intuition that models shouldn't require fine-tuning

Agenda

GRB number counts from BATSE & GBM

- *Selection effects* (truncation, censoring) — *obvious* (usually)
  Typically treated by "correcting" data
  Most sophisticated: product-limit estimators

- *"Scatter" effects* (measurement error, etc.) — *insidious*
  Typically ignored (average out? *No*—Eddington bias!)

Accounting for measurement error

Suppose $f(x|\theta)$ is a distribution for an observable, x (scalar or vector, $\vec{x} = (x, y, \ldots)$); and θ is unknown
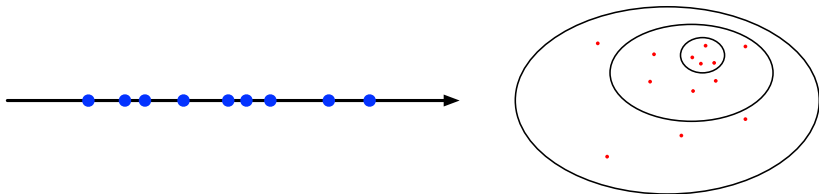
From N precisely measured samples, $\{x_i\}$, we can infer θ from

$$\mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta)$$
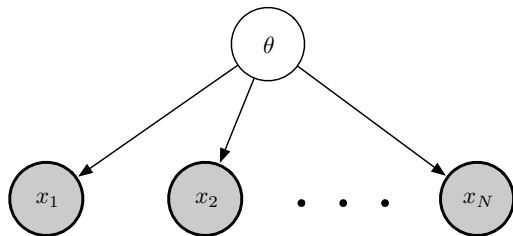
(A *binomial point process*)

$$p(\theta|\{x_i\}) \propto p(\theta)\mathcal{L}(\theta) = p(\theta, \{x_i\})$$

Posterior $\propto$ joint for params & data

Graphical representation

- Nodes/vertices = uncertain quantities (gray $\rightarrow$ known)
- Edges specify conditional dependence
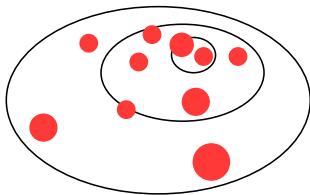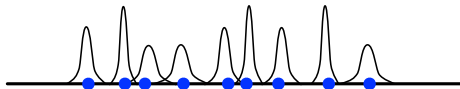- Absence of an edge denotes *conditional independence*

Graph specifies the form of the *joint distribution*:

$$p(\theta, \{x_i\}) = p(\theta)\, p(\{x_i\}|\theta) = p(\theta) \prod_i f(x_i|\theta)$$

Posterior from BT: $p(\theta|\{x_i\}) = p(\theta, \{x_i\})/p(\{x_i\})$

But what if the x data are *noisy*, $D_i = \{x_i + \epsilon_i\}$?

$\{x_i\}$ are now *uncertain (latent/hidden/incidental) parameters*

Member/item likelihoods quantify uncertainties: $\ell_i(x_i) = p(D_i|x_i)$

The joint PDF for *everything* is

$$
\begin{aligned}
p(\theta, \{x_i\}, \{D_i\}) &= p(\theta)\, p(\{x_i\}|\theta)\, p(\{D_i\}|\{x_i\}) \\
&= p(\theta) \prod_i f(x_i|\theta)\, \ell_i(x_i)
\end{aligned}
$$

The conditional (posterior) PDF for the unknowns is

$$
p(\theta, \{x_i\}|\{D_i\}) = \frac{p(\theta, \{x_i\}, \{D_i\})}{p(\{D_i\})} \propto p(\theta, \{x_i\}, \{D_i\})
$$

$$p(\theta, \{x_i\} | \{D_i\}) \;\; \propto \;\; p(\theta, \{x_i\}, \{D_i\})$$
$$= \;\; p(\theta) \prod_i f(x_i | \theta) \, \ell_i(x_i)$$

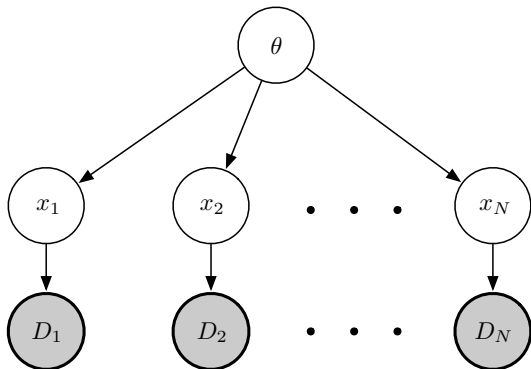*Marginalize over* $\{x_i\}$ to summarize inferences for θ

Marginalize over θ to summarize inferences for $\{x_i\}$

Profile likelihood (plugging in best-fit/MLE $\hat{x}_i$) is known to give (statistically) *inconsistent estimates* in this "density deconvolution" setting

Marginalizing gives sound estimates and implements shrinkage/borrowing strength

(See Loredo (2004) for tutorial examples)

Graphical representation

$$p(\theta, \{x_i\}, \{D_i\}) \;=\; p(\theta)\, p(\{x_i\}|\theta)\, p(\{D_i\}|\{x_i\})$$
$$=\; p(\theta)\prod_i f(x_i|\theta)\, p(D_i|x_i) \;=\; p(\theta)\prod_i f(x_i|\theta)\, \ell_i(x_i)$$

A two-level *hierarchical Bayes model*, *multi-level model* (MLM), or
probabilistic graphical model (PGM)

Agenda

Bayesian inference in a nutshell

Probability as generalized logic

   Probability quantifies the *strength of arguments*

   To appraise hypotheses, calculate probabilities for arguments
   from data and modeling assumptions to each hypothesis

   Use *all* of probability theory for this

Bayes's theorem

   $p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$

   Data *change* the support for a hypothesis $\propto$ ability of
   hypothesis to *predict* the observed data

Law of total probability

   $p(\text{Hypothes}\underline{\textbf{es}} \mid \text{Data}) = \sum p(\text{Hypothes}\underline{\textbf{is}} \mid \text{Data})$

   The support for a *composite* hypothesis must account for all
   the ways it could be true, via *marginalization*

Many roles for marginalization

Many composite hypotheses are of interest...

Credible regions

$$p(\theta \in \Delta | D, M) = \int_\Delta d\theta \; p(\theta | D, M)$$

Eliminate nuisance parameters

$$p(\phi | D, M) = \int d\eta \; p(\phi, \eta | D, M)$$

Model uncertainty & multi-model inference...

$$p(D | M_i) = \int d\theta_i \; p(\theta_i | M) \; \mathcal{L}(\theta_i)$$

Propagate uncertainty

Model has parameters θ; what can we infer about $F = f(\theta)$?

$$
\begin{aligned}
p(F|D, M) &= \int d\theta \, p(F, \theta|D, M) = \int d\theta \, p(\theta|D, M) \, p(F|\theta, M) \\
&= \int d\theta \, p(\theta|D, M) \, \delta[F - f(\theta)] \qquad [\textit{single-valued case}]
\end{aligned}
$$

Prediction

Given a model with parameters θ and present data D, predict future data D' (e.g., for *experimental design*):

$$
p(D'|D, M) = \int d\theta \, p(D', \theta|D, M) = \int d\theta \, p(\theta|D, M) \, p(D'|\theta, M)
$$

Hierarchical modeling (graphical models, multilevel models)

Learn population parameters by marginalizing over latent parameters for each member's actual (vs. measured) properties.

Learn a member's params by marginalizing over pop'n model and other members' params $\rightarrow$ *shrinkage* (beneficial bias!) — "the single most striking result of post-World War II statistical theory" (Efron 2010). Generalizes Malmquist/Eddington bias corrections.

Theme: Parameter space volume

Bayesian calculations sum/integrate over parameter/hypothesis space!

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space

- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters

- Model marginal likelihoods have parameter space volume factors that can penalize models for unncecessary complexity

- Prediction, uncertainty propagation, model averaging. . .

Many virtues of Bayesian methods can be attributed to this accounting for the "size" of parameter space. This idea does not arise naturally in frequentist statistics (but it can be added "by hand"—ignoring Fisher!).